Qian *et al. EURASIP Journal on Advances in Signal Processing* (2015) 2015:40
DOI 10.1186/s13634-015-0224-z

EURASIP Journal on Advances in Signal Processing
a SpringerOpen Journal

RESEARCH Open Access

Class-specific Gaussian-multinomial latent Dirichlet allocation for image annotation

Zhiming Qian*, Ping Zhong and Runsheng Wang

Abstract

Image annotation has been a challenging problem due to the well-known semantic gap between two heterogeneous information modalities, i.e., the visual modality referring to low-level visual features and the semantic modality referring to high-level human concepts. To bridge the semantic gap, we present an extension of latent Dirichlet allocation (LDA), denoted as class-specific Gaussian-multinomial latent Dirichlet allocation (csGM-LDA), in an effort to simulate the human's visual perception system. An analysis of previous supervised LDA models shows that the topics discovered by generative LDA models are driven by general image regularities rather than the semantic regularities for image annotation. To address this, csGM-LDA is introduced by using class supervision at the level of visual features for multimodal topic modeling. The csGM-LDA model combines the labeling strength of topic supervision with the flexibility of topic discovery, and the modeling problem can be effectively solved by a variational expectation-maximization (EM) algorithm. Moreover, as natural images usually generate an enormous size of high-dimensional data in annotation applications, an efficient descriptor based on Laplacian regularized uncorrelated tensor representation is proposed for explicitly exploiting the manifold structures in the high-order image space. Experimental results on two standard annotation datasets have shown the effectiveness of the proposed method by comparing with several state-of-the-art annotation methods.

Keywords: Image annotation; Latent Dirichlet allocation; Variational EM; Uncorrelated tensor representation; Laplacian regularization

1 Introduction

Automatic image annotation is a challenging work of tasks related to understanding what we see in a visual scene due to the well-known semantic gap [1]. Given an input image, the goal of image annotation is to assign meaningful tags to the image aiming at summarizing its visual contents. Such methods are becoming more and more important given the growing collections of both private and publicly available images. However, challenges for these methods often lie in three aspects: the inter-tag similarity problem that different tags may have similar visual contents, the tag locality problem that most tags are only related to their corresponding semantic regions, and the intra-tag diversity problem that the relevant regions for each tag at different images can be different.

The inter-tag similarity problem reveals the fact that the visual similarity does not always guarantee the semantic similarity, which in general is conflicting with the inherent assumption of many image annotation methods, e.g., some relevant methods [2,3] that perform tag propagations according to their visual similarities. To cope with this problem, it is emergent to develop more discriminative visual features that can be used to separate various visual contents for different tags. However, traditional vector representations in the form of bag-of-features or bag-of-words, such as the visual descriptor that quantizes SIFT local features [3] and the colored pattern appearance model (CPAM) [4], are usually incompetent for the intention. The reason is that these features usually ignore the high-order characteristics of natural images and might lead to the curse of dimensionality problem when requiring a relatively discriminative representation for describing the complex visual world. In practice, an image is intrinsically a two-dimensional or high-order tensor. To

*Correspondence: qianzhiming@nudt.edu.cn
College of Electronic Science and Engineering, National University of Defense Technology, Deya Road, 410073 Changsha, China

fairly evaluate the high-order characteristics of image contents, tensor representations [5,6], which can explicitly describe the multiple interrelated restrictions, might allow us to avoid the problem of curse of dimensionality.

To tackle the tag locality problem, one may employ local image features instead of holistic image features to describe the visual contents of a certain tag. The work in [7] considered each image as a bag of multiple segmented regions and predicted the tag of each region by a multiclass bag classifier. This method, however, heavily depends on the segmentation performance, which is very sensitive to the image noise. Recently, implicit image representations attract much attention on describing local regions. To reveal the tag locality, Bao et al. [8] introduced hidden concepts for decomposing holistic image representation into tag representations. Mesnil et al. [9] learned implicit representations for both the objects and their parts. Although these representations cannot explicitly describe the regions of a certain tag, they implicitly capture the tag's local visual contents by learning from large amount of annotated images. Thus, implicit image representation is nontrivial for tackling the tag locality problem in large-scale datasets.

Considering the problem of intra-tag diversity, a straightforward way is to set up the class-specific techniques [10,11] by treating annotation tags as class labels and learning the visual contents within each class. Although capable of identifying sets of visual contents discriminative for the classes of interest, these straightforward methods do not explicitly model the interclass and intraclass structures of visual distributions due to its lack of hierarchical content groupings. To facilitate the discovery of these structures, various hierarchical generative methods have been recently ported from the text to the vision literature. Among these methods, topic models, such as latent Dirichlet allocation (LDA) [12] and probabilistic latent semantic analysis (pLSA) [13], that consider probabilistic latent variable models for hierarchical learning have caused extensive interest. However, an analysis of previous supervised topic models [14] shows that the topics discovered by these models are driven by general image regularities rather than the semantic regularities for image annotation. For example, it has been noted in [14] that given a collection of movie reviews, LDA might discover topics as movie properties, such as genres, which are not central to the annotation task. Therefore, incorporating a class label variable into a generative model might tackle the intra-tag diversity problem well. Such extensions have been successfully applied into the classification task, such as class LDA (cLDA) [14], supervised LDA (sLDA) [15], class-specific-simplex LDA (css-LDA) [16], and so on.

In this paper, we develop a new extension of LDA coupled with Laplacian regularized uncorrelated tensor representation for learning semantics in the image data. Since tensor representation can well capture the high-order statistics and structures from the training data, the proposed representation method achieves an efficient compressed image representation by imposing noncorrelation constraints and Laplacian regularization in tensor factorization. Based on this representation, a three-level hierarchical probabilistic model, denoted as class-specific Gaussian-multinomial latent Dirichlet allocation (csGM-LDA), is developed by using class supervision at the level of visual features. In csGM-LDA, latent variables or topics are served as middle-level concepts for building the correspondences between visual features and annotation tags.

The core contributions of this paper are listed as follows:

- A novel hierarchical probabilistic model, namely csGM-LDA, is presented by combining the labeling strength of topic supervision with the flexibility of topic discovery, and can be effectively modeled by applying a variational EM algorithm.
- An effective image representation method, namely, Laplacian regularized uncorrelated tensor representation, is developed to explicitly consider the manifold structures in the high-order image space.
- By learning with csGM-LDA, a unified framework is introduced to infer the hierarchies of multiple modalities and predict tags for a new image. Benefiting from the exploration of hierarchical probabilistic inferences, the unified framework can be effectively conducted.

The rest of this paper is organized as follows. We first discuss the related work in Section 2. Then, we present Laplacian regularized uncorrelated tensor representation in Section 3. After that, the proposed model is described in Section 4. Moreover, quantitative experiments validating strong improvements by the proposed method are presented in Section 5. Finally, Section 6 draws the conclusion.

2 Related work

In this section, we outline research contributions which are most related to our work. We first review techniques for tensor-based image representation. Then, topic models are further discussed.

2.1 Tensor-based image representation

It is believed that the specialized structures of a visual object are intrinsically in the form of second or even higher order tensor [5]. To retain these high-order characteristics, tensors or multidimensional arrays become a natural choice for the visual representation. In practice, exact image representation as a full tensor is often redundant and impossible when coping with mass of images.

However, approximative image representation using tensor subspace learning techniques in many cases can be helpful for describing various visual objects. In this paper, we discuss two main kinds of tensor subspace learning (TSL) algorithms: supervised and unsupervised TSL.

Supervised TSL algorithms use concept-driven dimensionality reduction to achieve discriminant tensor subspaces by considering the subsequent classification or recognition tasks. This line of algorithms requires that either manual class labels or object priors in the training set can be applicable to a particular image classification [5,6] or object representation [17,18]. However, as image annotation system needs to handle a large number of classes and most classes may require many training samples due to significant intraclass shape and appearance variations, it is important that the learning does not involve any human interaction. This makes unsupervised TSL algorithms more appealing. Unsupervised TSL algorithms are actively explored for data-driven dimensionality reduction that uses low rank tensors to approximate the exact represented tensors. The extensions of principal component analysis (PCA) and singular value decomposition (SVD) are most familiar methods for the research on this line. By maximizing the variance measure, two-dimensional PCA (2DPCA) [19] represented an image by projecting it to principal components along the vertical direction of the image data. Then, generalized PCA (GPCA) [20] employed bilinear subspace analysis for dimensionality reduction with matrices. Later, the multilinear PCA (MPCA) [21] and uncorrelated MPCA (UMPCA) [22] were proposed for dimensionality reduction with tensors of any order. By minimizing the reconstruction error, the generalized low-rank approximation of matrices (GLRAM) [23] took into account the spatial correlation of image pixels within a localized neighborhood and applied bilinear transforms to the input image matrices. For higher-order tensors, the work in [24] used the high-order SVD (HOSVD) to decompose an ensemble of images into basis images that capture the different underlying factors of variations. Furthermore, concurrent subspaces analysis (CSA) [25] was presented as a generalization of GLRAM for higher-order tensors. Recently, multiple tensor rank-R decomposition (MTRD) [26] was proposed for approximating a higher-order tensor with a series of rank-R tensor approximations.

In this paper, we propose an unsupervised method with Laplacian regularized uncorrelated tensor representation to explicitly consider manifold structures in the high-order image space. That is, data points that are close in the intrinsic geometry of the image space shall thus be close to each other under the factorized tensor basis. By combining unsupervised TSL and Laplacian regularization, we can achieve a more discriminative descriptor which is much important for accurate semantic learning.

2.2 Topic models for image annotation

Topic models annotate images as the samples from a specific mixture of topics, where each topic is a distribution over image observations. Three alternatives of pLSA-based models that provided in [13] were presented by using asymmetric learning for semantic indexing of large image collections. Then, a Gaussian-multinomial pLSA (GM-pLSA) model [27] was presented to learn multimodal correlations from the image data by applying continuous feature vectors. Furthermore, the work in [28] extended pLSA to a higher-order formalism, so as to become applicable for more than two observable variables. However, pLSA-based models are incomplete in that they provide no probabilistic restriction on how to generate the training data. In these models, each image is represented as a list of the mixing proportions for topics, and there is no probabilistic inference for generating these numbers of topics. This leads to two problems: first, the number of modeling parameters grows linearly with the size of the training set, which leads to serious problems with overfitting; second, it is not clear how to assign probability to an image outside of the training set. To overcome these problems, it is much effective to endow the topic model with Dirichlet priors over topic parameters as they are conjugate to the multinomial distribution of the associated tags. The correspondence LDA (Corr-LDA) [29] was first presented for modeling the joint distribution of images and tags. To capture more general forms of association and allow the number of topics in the two data modalities to be different, topic regression multimodal latent Dirichlet allocation (tr-mmLDA) [30] was proposed by introducing a regression module to correlate the two sets of topics. Taking advantage of limited tagged training images and rich untagged images, the work in [31] proposed a regularized semi-supervised latent Dirichlet allocation (r-SSLDA) for learning visual concept classifiers in a semi-supervised way. However, several supervised methods [14-16] show that the topics discovered by LDA models are driven by general image regularities rather than the semantic regularities for image annotation. To address this, we propose a new three-level hierarchical probabilistic model by incorporating supervision into the extended LDA model, making the annotation applications be much effective than previous LDA models.

3 The proposed representation method

In this section, we first give the notations that are necessary in defining the multiway problem. Then, a tensor-based method is proposed for visual representation.

3.1 Notations and definitions

We follow the notation conventions in tensor algebra [32]. Vectors are usually denoted by lowercase letters, e.g., x; matrices by uppercase letters, e.g., X; and tensors by

calligraphic letters, e.g., $\mathcal{X}$. Their elements are denoted with indices in parentheses. Table 1 lists the key notations.

The inner product of two tensors of the same size $\mathcal{A}, \mathcal{B} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$ is defined as $\langle \mathcal{A}, \mathcal{B} \rangle = \sum_{i_1} \sum_{i_2} \cdots \sum_{i_N} \mathcal{A}(i_1, i_2, \cdots i_N) \cdot \mathcal{B}(i_1, i_2, \cdots i_N)$. Thus, the Frobenius norm of $\mathcal{A}$ can be denoted by $\|\mathcal{A}\|_F = \sqrt{\langle \mathcal{A}, \mathcal{A} \rangle}$. The n-mode matricization of $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$, denoted as $A_{(n)} \in \mathbb{R}^{I_n \times (I_1 \times \cdots \times I_{n-1} \times I_{n+1} \times \cdots \times I_N)}$, are obtained from $\mathcal{A}$ by varying the index i_n while keeping all the other indices fixed. The n-mode product of a tensor $\mathcal{A}$ by a matrix $U \in \mathbb{R}^{J_n \times I_n}$, denoted by $\mathcal{A} \times_n U$, is a tensor with entries:

$$(\mathcal{A} \times_n U)(i_1, \cdots, i_{n-1}, j_n, i_{n+1}, \cdots, i_N) = \sum_{i_n} \mathcal{A}(i_1, \cdots, i_N) \cdot U(j_n, i_n) \tag{1}$$

3.2 Laplacian regularized uncorrelated tensor representation

For image representation, we first represent a two-dimensional image I with an exact tensor representation, $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$, where $\{I_1, I_2\}$ denotes the size of the image and I_3 denotes the depth of image feature maps. In this representation, we consider the edge energies and the flow directions as supplementary to pixel-wise color information. At this point, we employ the Gaussian derivative (GD) and the difference of offset Gaussians (DoG), as defined in [33], for defining these features. The filter banks by convolving GD and DoG functions can be defined as:

$$\begin{cases} E_\sigma(I(x,y)) = \sqrt{|I(x,y) * \mathrm{GD}_{\sigma,0}(x,y)|^2 + |I(x,y) * \mathrm{GD}_{\sigma,\pi/2}(x,y)|^2} \\ F_\sigma(I(x,y)) = \sqrt{|I(x,y) * \mathrm{DoG}_{\sigma,0}(x,y)|^2 + |I(x,y) * \mathrm{DoG}_{\sigma,\pi/2}(x,y)|^2} \end{cases} \tag{2}$$

Table 1 List of key notations

Symbol	Description
$\mathcal{X}_n, \tilde{\mathcal{X}}_n, \mathcal{G}_n$	The representations of an original tensor, centered tensor, and its core tensor
$U^{(k)}, \tilde{U}_{(-k)}$	The k-mode transformed matrix and the Kronecker products except the matrix
$\tilde{W}, \tilde{D}, \tilde{L}$	The weight matrix of tensorial features, its diagonal matrix, and its Laplacian matrix
g_n, y_n	Vectorizations of the core tensor and the transformed tensor
$\alpha, \beta, \mu_c, \sigma_c$	Parameters of the Dirichlet distribution for latent topics, multinomial distribution for tags, mean and variance of Gaussian distribution for visual features
w_m, v_d	Symbol of the tag and the visual feature
y_m, z_d	Latent topics for the tag and the visual feature
W, V	Collections of tags and visual features
Y, Z	Collections of latent topics for tags and visual features

where (x, y) denotes the coordinates of the image and σ is a scale parameter. Then, the image can be represented by:

$$\mathcal{X} = [Y\ C_b\ C_r\ E_{\sigma/2}(Y)\ E_\sigma(Y)\ E_{2\sigma}(Y)\ F_{\sigma/2}(Y)\ F_\sigma(Y)\ F_{2\sigma}(Y)] \tag{3}$$

where Y, C_b, C_r are the three color channels obtained by transforming the original RGB image.

Let $\{\mathcal{X}_n | \mathcal{X}_n \in \mathbb{R}^{I_1 \times I_2 \times I_3}, n = 1, \cdots, N\}$ be a set of represented tensors from an image dataset. To learn uncorrelated features without loss of generality, training samples are subtracted to be zero-mean so that the constraint of uncorrelated features is the same as orthogonal projected features. Usually, the image space is always of high dimensionality. However, image contents are typically embedded in a lower dimensional tensor subspace, in analogy to the dimensional reduction problem that considers both feature selection (i.e., give a more informative description to pixels) and spatial correlation. Thus, the task is to find a tensor subspace that captures most of the characteristics in the input space, i.e., define multilinear transformations $\{U^{(k)} \in \mathbb{R}^{I_k \times J_k}, k = 1, 2, 3 | \left(U^{(k)}\right)^T U^{(k)} = I\}$ that rewrite the original tensor as:

$$\tilde{\mathcal{X}}_n \approx \mathcal{G}_n \times_1 U^{(1)} \times_2 U^{(2)} \times_3 U^{(3)} \tag{4}$$

where $\tilde{\mathcal{X}}_n = \mathcal{X}_n - \bar{\mathcal{X}}$, $\bar{\mathcal{X}} = (1/N) \sum_{n=1}^N \mathcal{X}_n$ and $\mathcal{G}_n \in \mathbb{R}^{J_1 \times J_2 \times J_3}$ is a core tensor. To capture the core tensor, we determine the objective function with the following minimization problem:

$$\left\{ U^{(k)} \in \mathbb{R}^{I_k \times J_k}, \mathcal{G}_n | \left(U^{(k)}\right)^T U^{(k)} = I \right\} = \arg \min_{U^{(k)}, \mathcal{G}_n} \Psi + \lambda \mathcal{M}_\mathcal{G} \tag{5}$$

where $\Psi = \sum_{n=1}^N \left\| \tilde{\mathcal{X}}_n - \mathcal{G}_n \times_1 U^{(1)} \times_2 U^{(2)} \times_3 U^{(3)} \right\|_F^2$, and $\mathcal{M}_G = G^T \tilde{L} G = \sum_{ij} \tilde{W}(i,j) \left\| g_i - g_j \right\|_2^2$. Here, $\left\{ g_n \in \mathbb{R}^D, D = \prod_{k=1}^3 J_k \right\}$ is the vectorization of $\mathcal{G}_n$, and G is a matrix with the column of g_n. Define the diagonal matrix $\tilde{D}$ whose entries are column sums of the weight matrix $\tilde{W}$, and $\tilde{L} = \tilde{D} - \tilde{W}$ is a Laplacian matrix. The nearest neighbor graph is used to construct the weight matrix by finding the nearest neighbors for each image data. We use $NN(\mathcal{X}_n)$ to denote the set of K_{NN} nearest neighbors of $\mathcal{X}_n$. The weight matrix can be simply defined as:

$$W(i,j) = \begin{cases} 1 & \text{if } \mathcal{X}_i \in NN(\mathcal{X}_j) \text{ or } \mathcal{X}_j \in NN(\mathcal{X}_i) \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

To solve the problem defined in Equation 5, an alternating iteration scheme is applied. Given all the projection matrices $\{U^{(k)} \in \mathbb{R}^{I_k \times J_k}, k = 1,2,3\}$, we can obtain that:

$$\{\mathcal{G}_n, n = 1, \cdots, N\} = \arg\min_{\mathcal{G}_n} \Psi_{\mathcal{Y}} + \lambda \mathcal{M}_{\mathcal{G}} \tag{7}$$

where $\Psi_{\mathcal{Y}} = \sum_{n=1}^{N} \|y_n - g_n\|_F^2$, $y_n = vec\left(\tilde{\mathcal{X}}_n \times_1 \left(U^{(1)}\right)^T \times_2 \left(U^{(2)}\right)^T \times_3 \left(U^{(3)}\right)^T\right)$. The above function defines a quadratic programming problem and can be solved linearly by using the Newton-Raphson method [12] with the following iteration:

$$g_n^{(t+1)} = g_n^{(t)} - (I + \lambda \tilde{L})^{-1} \left(g_n^{(t)} - y_n + \lambda \sum_j \tilde{L}(n,j) g_j^{(t)} \right) \tag{8}$$

Given the core tensor $\mathcal{G}_n$, we can rewrite Equation 5 as:

$$\{U^{(k)} \in \mathbb{R}^{I_k \times J_k}, k = 1,2,3 | \left(U^{(k)}\right)^T U^{(k)} = I\} = \arg\min_{U^{(k)}} \Psi \tag{9}$$

Since the projection to a high-order tensor subspace consists of several projections to the corresponding vector subspaces, the optimization can be iteratively solved by finding the k-mode projection that maximizes the scatter in the k-mode vector subspace. To optimize $U^{(k)}$, we first define two scatter matrices:

$$\begin{cases} \Phi_G^{(k)} = \sum_{n=1}^{N} G_{n(k)} \tilde{U}_{(-k)}^T \tilde{U}_{(-k)} G_{n(k)}^T \\ \Phi_X^{(k)} = \sum_{n=1}^{N} \tilde{X}_{n(k)} \tilde{U}_{(-k)} G_{n(k)}^T \end{cases} \tag{10}$$

where $\tilde{U}_{(-k)} = U^{(1)} \otimes \cdots \otimes U^{(k-1)} \otimes U^{(k+1)} \otimes \cdots \otimes U^{(3)}$ and '$\otimes$' denotes the *Kronecker product*. Then, the solution to Equation 9 can be achieved by:

$$U^{(k)} = \left(\Phi_G^{(k)}\right)^{-1} \Phi_X^{(k)} \tag{11}$$

The pseudo code for the proposed representation method is described in Algorithm 1. For this representation, a full solution is referring to the formalism in Equation 5. However, the alternating solution for this problem is quadratic with respect to the number of the image dataset, which is much expensive for image representation when dealing with a large dataset. In real applications, we perform the above representation method with a much smaller size by first using graph shift [34] for image clustering and then learning the representation for each group. Noticeably, the image data of one group should subtract the projections of previous multilinear transformations to preserve the orthogonality.

Algorithm 1: Laplacian regularized uncorrelated tensor representation

Input: A group of image samples $\{\mathcal{X}_n | \mathcal{X}_n \in \mathbb{R}^{I_1 \times I_2 \times I_3}, n = 1, \cdots, N\}$

Output: Multilinear transformations $\{U^{(k)} \in \mathbb{R}^{I_k \times J_k}, k = 1,2,3\}$ and core tensors $\{\mathcal{G}_n, n = 1, \cdots, N\}$

Step 1: Center the input image samples as $\{\tilde{\mathcal{X}}_n | \tilde{\mathcal{X}}_n = \mathcal{X}_n - \bar{\mathcal{X}}, n = 1, \cdots, N\}$.

Step 2: Calculate the eigen-decomposition of $S^{(k)} = \sum_{n=1}^{N} \tilde{X}_{n(k)} \tilde{X}_{n(k)}^T$, and set $U^{(k)}$ to consist of the least number of the eigenvectors corresponding to the largest J_k eigenvalues.

Step 3: Iteratively compute $\{\mathcal{G}_n, n = 1, \cdots, N\}$ according to Equation 8.

Step 4: Update $\{U^{(k)} \in \mathbb{R}^{I_k \times J_k}, k = 1,2,3\}$ according to Equation 11.

Step 5: Alternately perform steps 3 and 4 until convergence.

4 The proposed annotation method

In this section, we first describe the proposed method for image annotation. Then, we turn our attention to parameter estimation for the modeling problem. Finally, a unified framework is presented to infer the hierarchies of multiple modalities and predict tags for a new image.

4.1 Class-specific Gaussian-multinomial latent Dirichlet allocation

The proposed model, csGM-LDA, is a supervised probabilistic model for learning multiple relationships between images and tags. The basic idea is that the visual and semantic modalities of images are represented as random mixtures over multiple latent topics, where each topic is characterized by a distribution over each modality. In Figure 1, the generative process of csGM-LDA for an image-tag pair with M associated tags and D visual features is given as follows:

1. Draw an image topic proportion $\theta \sim P_\Pi(\theta; \alpha)$
2. For each associated tag $w_m \in \mathcal{W} = \{1, \cdots, C\}, m = 1, \cdots, M$

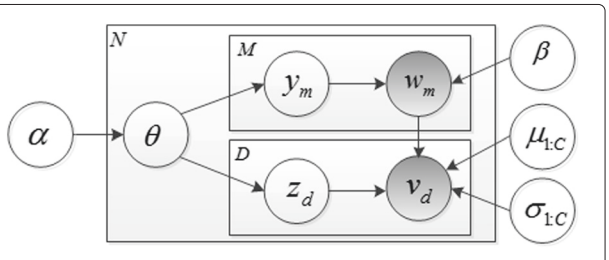

Figure 1 Illustration of csGM-LDA.

(a) Draw a topic assignment, $y_m \sim P_{Y|\Pi}(y_m|\theta), y_m \in \mathcal{T}_y = \{1, \cdots, K\}$
(b) Draw a tag, $w_m \sim P_{W|Y}(w_m|y_m; \beta), w_m \in \mathcal{W} = \{1, \cdots, C\}$

3. For each visual word $v_d, d = 1, \cdots, D$

(a) Draw a topic assignment, $z_d \sim P_{Z|\Pi}(z_d|\theta), z_d \in \mathcal{T}_z = \{1, \cdots, K\}$
(b) Draw a visual description, $v_d \sim P_{V|Z,W}(v_d|z_d, w_m; \mu_{w_m}, \sigma_{w_m})$

Similar to the earlier LDA extensions, $P_\Pi(\cdot)$ is a Dirichlet distribution on the topic simplex $\theta \in \mathbb{R}^K$ with the parameter $\alpha \in \mathbb{R}^K$, $P_{Y|\Pi}(\cdot)$ and $P_{Z|\Pi}(\cdot)$ are two multinomial distributions over the topic simplex θ, $P_{W|Y}(\cdot)$ is a categorical distribution over a topic y_m with the parameter $\beta \in \mathbb{R}^{K\times C}$ where $\beta(k,c) = p(w_m = c|y_m = k)$, and $P_{V|Z,W}(\cdot)$ is a Gaussian distribution over a topic z_d with the class-dependent parameters $\{\mu_{w_m}(z_d, d), \sigma_{w_m}(z_d, d)|\mu_{w_m}, \sigma_{w_m} \in \mathbb{R}^{K\times D}\}$. In this way, semantic topics generate tags or classes for images with each defining a prior distribution in visual topic space. Taking Bayesian rules, the joint distribution of $\{\theta, y_m, w_m, z_d, v_d\}$ for each image is given by:

$$\begin{aligned} &P_{\Pi,Y,W,Z,V}(\theta, y_m, w_m, z_d, v_d|\alpha, \beta, \mu_{w_m}, \sigma_{w_m}) = \\ &P_\Pi(\theta|\alpha)P_{Y|\Pi}(y_m|\theta)P_{W|Y}(w_m|y_m, \beta)P_{Z|\Pi}(z_d|\theta) \\ &P_{V|Z,W}(v_d|z_d, \mu_{w_m}, \sigma_{w_m}) \end{aligned} \tag{12}$$

The modeling problem is then to maximize the log likelihood of the following marginal function:

$$\begin{aligned} &P_{W,V}(W, V|\alpha, \beta, \mu_{1:C}, \sigma_{1:C}) = \int_{\theta_n} \prod_{n=1}^{N} \prod_{m=1}^{M} \prod_{d=1}^{D} \sum_{y_{nm}} \sum_{z_{nd}} \\ &P_{\Pi,Y,W,Z,V}(\theta_n, y_{nm}, w_{nm}, z_{nd}, v_{nd}|\alpha, \beta, \mu_{w_{nm}}, \sigma_{w_{nm}}) d\theta_n \end{aligned} \tag{13}$$

In csGM-LDA, the parameters $\{\alpha, \beta, \mu_{1:C}, \sigma_{1:C}\}$ are dataset-level parameters, assumed to be sampled once in the process of generating a set of images. The variables θ_n is an image-level variable, sampled once per image. The variables y_{nm} and w_{nm} are tag-level variables, sampled once for each annotated tag. And the variables z_{nd} and v_{nd} are structure-level variables, sampled once for each visual description. Structural models similar to that shown in Figure 1 are often studied in Bayesian statistical modeling, where they are referred to as conditionally independent hierarchical models. Indeed, as we discuss in the following subsection, we adopt the empirical Bayes approach to estimating parameters with a variational EM algorithm.

4.2 Parameter estimation via variational inference

In this section, we describe a convexity-based variational algorithm for parameter estimation. The basic idea of convexity-based variational inference is to make use of Jensen's inequality to obtain an adjustable lower bound on the log likelihood [13]. Essentially, a family of lower bounds is usually indexed by a set of variational parameters. The variational parameters are chosen by an optimization procedure that attempts to find the tightest possible lower bound. In particular, the objective function in Equation 13 is usually intractable due to the couplings between θ_n and $\{\beta, \mu_{1:C}, \sigma_{1:C}\}$ in the summation over latent topics. By dropping these edges and endowing the simplified graphical model with free variational parameters, we obtain a new family of distributions on the latent variables as seen in Figure 2. The variational distribution $q(\theta, Y, Z|\eta, \phi, \zeta)$ can be characterized by:

$$q(\theta, Y, Z|\eta, \phi, \zeta) = \prod_{n=1}^{N} q(\theta_n|\eta_n) \prod_{m=1}^{M} q(y_{nm}|\phi_{nm}) \prod_{d=1}^{D} q(z_{nd}|\zeta_{nd}) \tag{14}$$

where $\{Y, Z\}$ are the collections of latent topics, the Dirichlet parameter η_n, and the multinomial parameters $\{\phi_{nm}, \zeta_{nd}\}$ are the free variational parameters.

Having specified a simplified family of probability distributions, the next step is to set up an optimization problem that determines the values of the variational parameters:

$$\begin{aligned} \{\eta^*, \phi^*, \zeta^*\} = \arg \min_{\{\eta,\phi,\zeta\}} &D_{KL}\left(q(\theta, Y, Z|\eta, \phi, \zeta)\right. \\ &\left.||P_{\Pi,Y,Z}(\theta, Y, Z|W, V, \alpha, \beta, \mu_{1:C}, \sigma_{1:C})\right) \end{aligned} \tag{15}$$

where $D_{KL}(\cdot)$ is the Kullback-Leibler (KL) divergence and $\{W, V\}$ are the corresponding collections of their lowercase variables. To achieve this minimization, we begin with the expression of the true log-likelihood for an image-tag pair $\{W, V\}$ by bounding the log likelihood of $P_{W,V}(W, V|\alpha, \beta, \mu_{1:C}, \sigma_{1:C})$ using the Jensen's inequality:

$$\begin{aligned} &\log P_{W,V}(W, V|\alpha, \beta, \mu_{1:C}, \sigma_{1:C}) \geq \\ &E_q[\log P_{\Pi,Y,W,Z,V}(\theta, Y, W, Z, V|\alpha, \beta, \mu_{1:C}, \sigma_{1:C})] \\ &-E_q[\log q(\theta, Y, Z|\eta, \phi, \zeta)] \end{aligned} \tag{16}$$

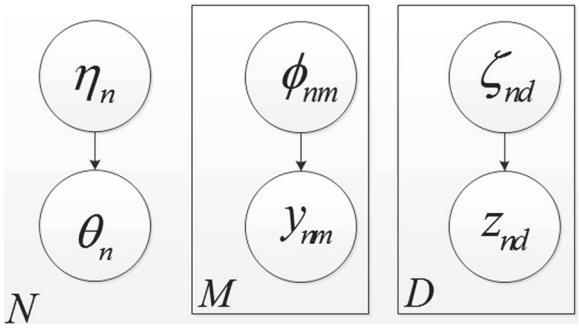

Figure 2 The variational model for approximating csGM-LDA.

It can be easily verified that the difference between the two sides of the above inequation is the KL divergence that provided in Equation 15. Let $\mathcal{L}(\eta,\phi,\zeta;\alpha,\beta,\mu_{1:C},\sigma_{1:C})$ denote the right-hand side of the inequation, we have:

$$\begin{aligned}&\log P_{W,V}(W,V|\alpha,\beta,\mu_{1:C},\sigma_{1:C}) = \mathcal{L}(\eta,\phi,\zeta;\alpha,\beta,\mu_{1:C},\sigma_{1:C}) + \\ &D_{KL}\left(q(\theta,Y,Z|\eta,\phi,\zeta)||P_{\Pi,Y,Z}(\theta,Y,Z|W,V,\alpha,\beta,\mu_{1:C},\sigma_{1:C})\right)\end{aligned} \tag{17}$$

This shows that maximizing the lower bound $\mathcal{L}(\eta,\phi,\zeta;\alpha,\beta,\mu_{1:C},\sigma_{1:C})$ with respect to $\{\eta,\phi,\zeta\}$ is equivalent to minimizing the above KL divergence. As seen in the Appendix, we obtain pair of updates:

$$\begin{cases}\phi_{nm}(k) \propto \beta(k,w_{nm}) \exp\left(\Psi(\eta_n(k)) - \Psi\left(\sum_{j=1}^{K}\eta_n(j)\right)\right) \\ \zeta_{nd}(k) \propto \left(\prod_{m=1}^{M} P_{V|Z,W}(v_{nd}|k,\mu_{w_{nm}},\sigma_{w_{nm}})\right) \\ \quad \exp\left(\Psi(\eta_n(k)) - \Psi\left(\sum_{j=1}^{K}\eta_n(j)\right)\right) \\ \eta_n(k) = \alpha(k) + \sum_{m=1}^{M}\phi_{nm}(k) + \sum_{d=1}^{D}\zeta_{nd}(k)\end{cases} \tag{18}$$

Based on the above variational inference, parameter estimation with respect to $\{\alpha,\beta,\mu_{1:C},\sigma_{1:C}\}$ yields the following EM algorithm:

1. *E-step:* For each image, find the optimizing values of the variational parameters$\{\eta,\phi,\zeta\}$.
2. *M-step*: Maximize the resulting lower bound $\mathcal{L}(\eta,\phi,\zeta;\alpha,\beta,\mu_{1:C},\sigma_{1:C})$ on the log likelihood of $P_{W,V}(W,V|\alpha,\beta,\mu_{1:C},\sigma_{1:C})$ with respect to the model parameters$\{\alpha,\beta,\mu_{1:C},\sigma_{1:C}\}$.

The update in the M-step corresponds to finding maximum likelihood estimates with expected sufficient statistics for each image under the approximate posterior which is computed by the variational parameters. Thus, we can also maximize the lower bound $\mathcal{L}(\eta,\phi,\zeta;\alpha,\beta,\mu_{1:C},\sigma_{1:C})$ with respect to the parameters $\{\alpha,\beta,\mu_{1:C},\sigma_{1:C}\}$. In Appendix, we show that the update in the M-step for the Dirichlet parameter α can be implemented with an efficient Newton-Raphson method. The parameters $\{\beta,\mu_{1:C},\sigma_{1:C}\}$ can be obtained as:

$$\begin{cases}\beta(k,c) \propto \sum_{n=1}^{N}\sum_{m=1}^{M}\phi_{nm}(k)\delta(c,w_{nm}) \\ \mu_c(k,d) = \frac{\sum_{n=1}^{N}\sum_{m=1}^{M}\zeta_{nd}(k)\delta(c,w_{nm})v_{nd}}{\sum_{n=1}^{N}\sum_{m=1}^{M}\zeta_{nd}(k)\delta(c,w_{nm})} \\ \sigma_c^2(k,d) = \frac{\sum_{n=1}^{N}\sum_{m=1}^{M}\zeta_{nd}(k)\delta(c,w_{nm})(v_{nd}-\mu_c(k,d))^2}{\sum_{n=1}^{N}\sum_{m=1}^{M}\zeta_{nd}(k)\delta(c,w_{nm})}\end{cases} \tag{19}$$

We summarize the parameter estimation algorithm in Algorithm 2. This is a standard EM process, and the lower bound $\mathcal{L}(\eta,\phi,\zeta;\alpha,\beta,\mu_{1:C},\sigma_{1:C})$ is a concave function. Therefore, Algorithm 2 is convergent. From the pseudo code, it is clear that each iteration of the E-step for csGM-LDA requires $\mathcal{O}(NMDK)$ operations. Empirically, we find that the number of iterations in this step is in proportion to the number of tags and the dimensionality of visual features. Parameter estimation in M-step for $\{\beta,\mu_{1:C},\sigma_{1:C}\}$ requires $\mathcal{O}(NCMDK)$ operations. And the number of iterations required for the Newton-Raphson method is linear to the dimensionality of α. Therefore, each EM iteration yields a total number of operations roughly on $\mathcal{O}(NM(C+M+D)DK)$. The number of iterations for the EM algorithm is mainly determined by the number of involved parameters, which is in proportion to $C(1+D)K$. Thus, the complexity for building the proposed model is about $\mathcal{O}\left(NCM(C+M+D)D^2K^2\right)$. When coping with large-scale data (i.e., $N \gg K,C,D$), the complexity of our modeling system is approximately linear to the number of images, which is much effective by comparing with the typical quadratic annotation models (e.g., pLSA [13] that requires $\mathcal{O}\left(N^2KC\right)$ operations, GM-pLSA [27] that yields the number of operations roughly on $\mathcal{O}\left(N^2K^2C\right)$, and so on).

Algorithm 2: Parameter estimation for csGM-LDA
Input: Observations $\{v_{nd},w_{nm},n=1,\cdots,N,m=1,\cdots,M,d=1,\cdots,D\}$
Output: Modeling parameters $\{\alpha,\beta,\mu_{1:C},\sigma_{1:C}\}$

repeat
In E-step:
For each image in the image dataset, initialize $\phi_{nm}(k)=\zeta_{nd}(k)=1/K$ by indexing all m, d, and k;
Initialize $\eta_n(k)=\alpha(k)+M/K$ by indexing all k.
for n = 1:N
repeat
Update $\{\eta,\phi,\zeta\}$ according to Equation 18.
until convergence
In M-step:
Initialize $\alpha(k)=1/K$ for all k. Update α with:
$\alpha^{(i+1)}(k)=\alpha^{(i)}(k)-$
$\left(\frac{\partial\mathcal{L}}{\partial\alpha^{(i)}(k)} - \frac{\sum_{j=1}^{K}\frac{\partial\mathcal{L}}{\partial\alpha^{(i)}(j)}\Big/\left(N\Psi'(\alpha^{(i)}(j))\right)}{1\Big/\left(N\Psi'\left(\sum_{j=1}^{K}\alpha^{(i)}(j)\right)\right)+\sum_{j=1}^{K}1/\left(N\Psi'(\alpha^{(i)}(j))\right)}\right)\Big/\left(N\Psi'(\alpha^{(i)}(k))\right)$
Update $\{\beta,\mu_{1:C},\sigma_{1:C}\}$ according to (19).
until convergence

4.3 A unified framework for image annotation

The unified framework is illustrated in Figure 3. Laplacian regularized uncorrelated tensor representation is first performed on the whole dataset. Then, both the semantic and the visual modalities are incorporated into building the proposed model. To annotate a new image, we first get the corresponding core tensor and then predict potential tags with Bayesian inference.

Given an observed visual image, our goal for image annotation is to estimate the posterior distribution of the

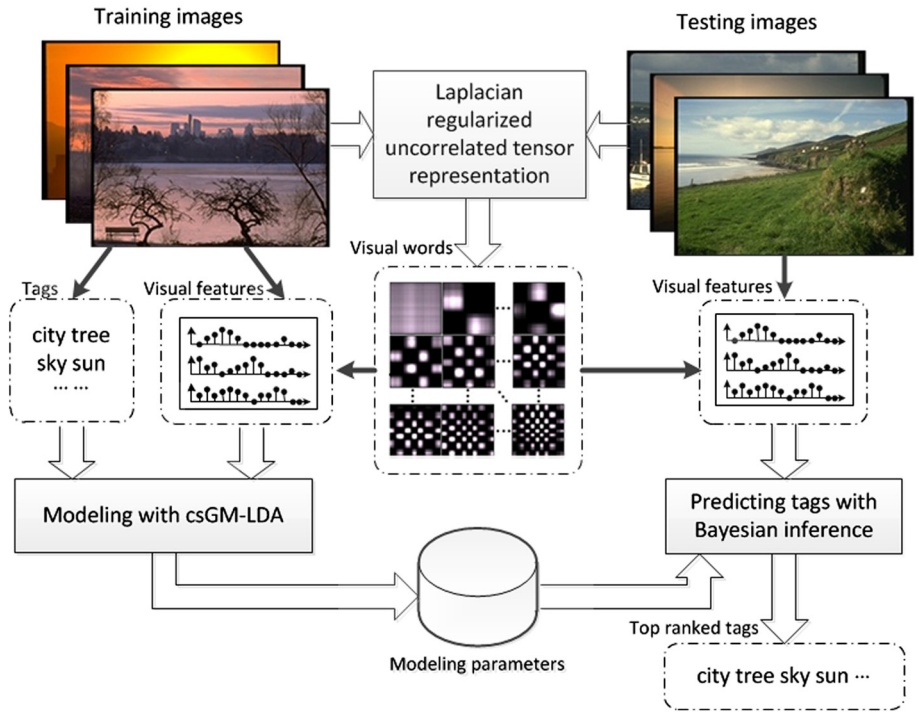

Figure 3 Illustration of the proposed framework for image annotation.

annotated tags. Taking Bayesian rules, we can infer their posterior probability by:

$$P_{W|V}(w|v_{1:D},\alpha,\beta,\mu_w,\sigma_w) = \frac{P_{W,V}(w,v_{1:D}|\alpha,\beta,\mu_w,\sigma_w)}{P_V(v_{1:D}|\alpha,\beta,\mu_w,\sigma_w)} = \frac{\int_\theta P_\Pi(\theta|\alpha)\left(\sum_y P_{Y|\Pi}(y|\theta)P_{W|Y}(w|y,\beta)\right)\left(\sum_{z_d} P_{Z|\Pi}(z_d|\theta)P_{V|Z,W}(v_d|z_d,\mu_w(z_d,d),\sigma_w(z_d,d))\right)d\theta}{\int_\theta P_\Pi(\theta|\alpha)\left(\sum_{z_d} P_{Z|\Pi}(z_d|\theta)P_{V|Z,W}(v_d|z_d,\mu_w(z_d,d),\sigma_w(z_d,d))\right)d\theta} \tag{20}$$

For simplicity, we employ Monte Carlo inference [35] to approximate the integral of θ. To get the above probability, we first generate samples $\{\theta^s, s = 1, \cdots, S\}$ from the posterior $\theta^s \sim P_\Pi(\theta|\alpha)$. Then, the tags' probability can be rewritten by:

$$P_{W|V}(w|v_{1:D},\alpha,\beta,\mu_w,\sigma_w) \approx \frac{\sum_{s=1}^{S}\left(\sum_y P_{Y|\Pi}(y|\theta^s)P_{W|Y}(w|y,\beta)\right)\left(\sum_{z_d} P_{Z|\Pi}(z_d|\theta^s)P_{V|Z,W}(v_d|z_d,\mu_w(z_d,d),\sigma_w(z_d,d))\right)}{\sum_{s=1}^{S}\sum_{z_d} P_{Z|\Pi}(z_d|\theta^s)P_{V|Z,W}(v_d|z_d,\mu_w(z_d,d),\sigma_w(z_d,d))} \tag{21}$$

5 Experiments

To evaluate the performance of our annotation framework, we set up several quantitative experiments. First, we investigate the effects of the setting parameters by conducting cross validation to select the best parameters for our proposed model. Then, we give a comparison of different image representations and validate the effectiveness of our representation method. Finally, we evaluate the proposed method on two benchmark datasets and report the results over state of the art.

5.1 Datasets and representations

We evaluate the proposed framework on two well-known benchmarks: Corel-5K [36] and ESP-Game [37]. The details of the two image datasets are shown in Table 2. The training percentage of each dataset is set as 80%, a validation set occupies 10% of the total images, and the remainder is the test set.

To get a reasonable size that keeps the images from serious deterioration for our representation method, we fix the size of images in Corel-5K and ESP-Game as 128×192 and 225×169, respectively. The vectorization of the core tensor constituting the D-dimensional Laplacian regularized uncorrelated tensorial vector (LGUTV) can be viewed as an image descriptor, with each item corresponding to an uncorrelated elementary multilinear projection. In our experiment, the dimensionality of LGUTV is fixed as 128 for each group in both the two datasets. We further divide the Corel-5K and ESP-Game into five and ten groups by graph shift [34], resulting in a 640-dimensional vector and a 1,280-dimensional vector for these two datasets, respectively. In addition, we

Table 2 Statistics of two image datasets

Dataset	Number of tags	Number of images	Tags per image
Corel-5K	5,000	374	3.4
ESP-Game	20,000	268	4.7

compare the proposed representation method with several common representation methods, i.e., the quantified color histograms with 16 bins in each color channel for RGB, LAB, HSV representations (C-HIST), the quantified SIFT features both extracted densely on a multiscale grid (D-SIFT) or for Harris-Laplacian interest points (H-SIFT) [3], local binary patterns (LBP) [38], and CPAM [4]. To get a proper evaluation of these image descriptors, we set their dimensions equal to that of LGUTV.

5.2 Evaluation criteria and baselines

The performance of image annotation is evaluated by comparing the automatically generated tags for the test set with the human-produced ground truth. In this paper, we give the following two measures for annotation evaluation. Firstly, F_1 score is measured by computing precision (*Prec*) and recall (*Rec*) for fixed annotation length with the five most relevant tags.

$$F_1 = \frac{2 \times Prec \times Rec}{Prec + Rec} \tag{22}$$

Note that each image is forced to be annotated with five tags, although the image might have fewer or more tags in the ground truth. Therefore, even if a model predicts all ground truth tags with a significantly higher probability than other tags, we will not measure perfect precision and recall. Thus, we also measure the precision at different levels of the recall for assessing the general annotation performance. The mean average precision (*mAP*) [13] over tags are found by computing for each tag the average of precisions measured after each relevant image is retrieved.

$$mAP = \frac{1}{N_q} \sum_{q=1}^{N_q} \sum_{i \in rel(q)} \mathrm{Prec}(i) \,/\, \big|rel(q)\big| \tag{23}$$

where Prec(i) is the precision of the correctly retrieved images at rank i in the ranking results of a query q, $rel(q)$ is the set of relevant images for this query, and N_q is the number of all queries.

For all images in the two standard datasets, our methods are compared with several most relevant and state-of-the-art methods, including TagProp [3], pLSA [13], GM-pLSA [27], GM-LDA [29], Corr-LDA [29], topic regression multimodal latent Dirichlet allocation (tr-mmLDA) [30], and css-LDA [16].

5.3 Investigate the impact of the setting parameters

We first investigate the neighborhood size K_{NN} for the graph construction and the tradeoff parameter λ for evaluating the effectiveness of Laplacian regularization. These two parameters reflect different facets of data construction; we joint discuss the sensitivity of these two parameters. As seen in Figure 4, we measure the F_1 scores for different parameters by setting the number of latent topics equal to the number of tags. Then, we choose to set the two parameters as the most promising for the two datasets in the following experiments, which are $K_{NN} = 10, \lambda = 0.1$ for Corel-5K and $K_{NN} = 15, \lambda = 0.1$ for ESP-Game, respectively. Besides, we observe that Laplacian regularization can achieve a more effective representation for better semantic learning by comparing the case that $\lambda \neq 0$ with the case that $\lambda = 0$.

Regarding topic models, they require the number of latent topics (i.e.,K) to be estimated, as this hyper parameter defines the capacity of the model. We analyze the parameter with a cross validation scheme. As seen in Figure 5, the improvement of annotation performance grows much slowly when the number of latent topics arrives at 100 for both the two datasets. When this number increases over the total number of tag vocabularies, csGM-LDA might suffer from the overfitting problem.

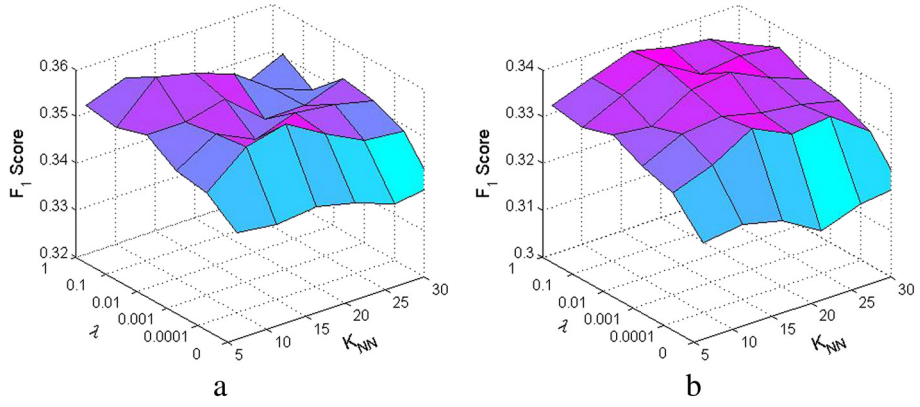

Figure 4 Investigation of the neighborhood size and the tradeoff parameter. F_1 score is measured to evaluate the performance of csGM-LDA by varying the two parameters on **(a)** Corel-5K and **(b)** ESP-Game.

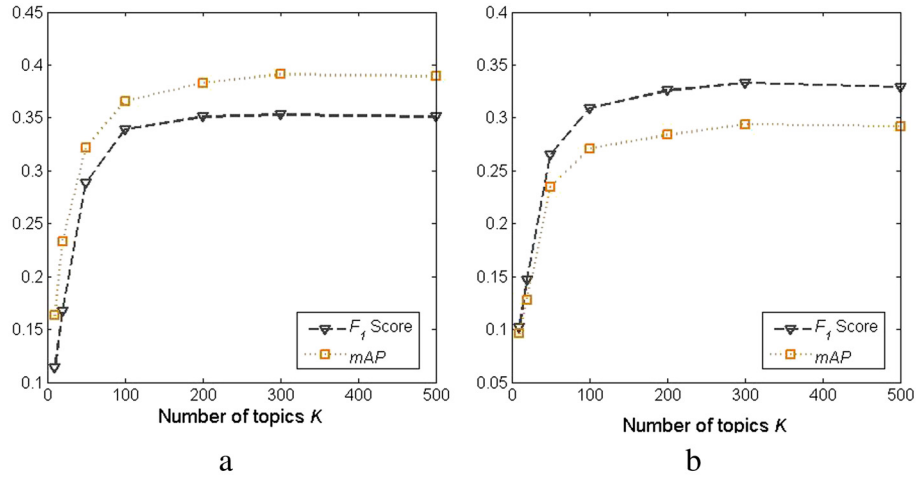

Figure 5 Investigation of the number of latent topics. The performance of csGM-LDA is evaluated by taking different number of latent topics in terms of F_1 score and *mAP* on **(a)** Corel-5K and **(b)** ESP-Game.

For example, the measures of F_1 score and *mAP* in both Figure 5a,b decrease when the number of latent topics is larger than 300. In the following experiments, we set the parameter as $K = 300$, which achieves the best performance for both the two datasets.

5.4 Evaluation of different image representations

In this paper, we argue that traditional vector-based image representations ignore high-order characteristics in image space, and thus, we combine unsupervised TSL and Laplacian regularization for achieving a more discriminative descriptor, i.e., LGUTV. To evaluate its effectiveness, we compare this descriptor with several state-of-the-art image descriptors on the two datasets by measuring F_1 Scores for the results of csGM-LDA. As seen in Figure 6, the performance of modeling csGM-LDA with LGUTV achieves the best performance by comparing with others on both the two datasets, confirming that tensor representation is most likely to provide a discriminative descriptor for recognizing the complex visual scenes.

5.5 Comparison with existing methods

In this section, we perform the annotation tasks on both the Corel-5K and ESP-Game datasets by comparing the proposed method with others. In Table 3, we report the performance by measuring both F_1 score and *mAP* for different methods based on LGUTV. On both two datasets, we observe that class-specific methods (e.g., TagProp, css-LDA, and csGM-LDA) obviously perform better than others. This is consistent with the analysis in the introduction since the generative domains of the complex visual world are much hard to describe. In addition, the measures for ESP-Game are consistently lower than that for Corel-5K because the retrieval tasks on which these measures are now computed are more challenging: an average of nearly 2,000 images for testing (versus 500 images) are ranked.

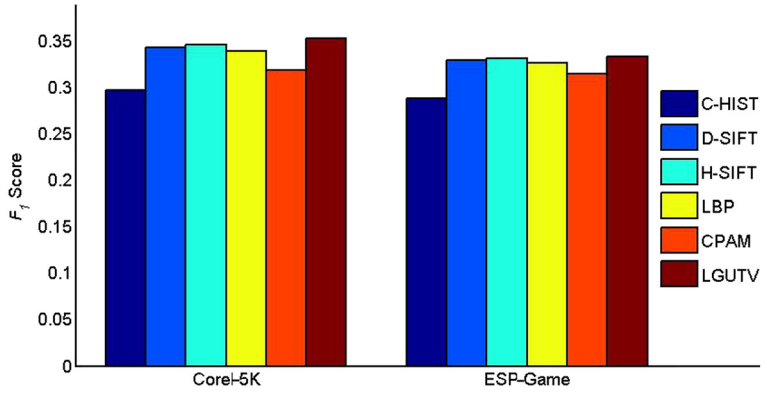

Figure 6 Annotation performance of csGM-LDA with different image representations.

Table 3 Comparison of different methods on both Corel-5K and ESP-Game

	Corel-5K		ESP-game	
Methods	F_1 **Score**	***mAP***	F_1 **score**	***mAP***
TagProp	0.342	0.374	0.323	0.281
pLSA	0.193	0.235	0.185	0.156
GM-pLSA	0.254	0.289	0.213	0.179
GM-LDA	0.276	0.303	0.224	0.192
Corr-LDA	0.316	0.354	0.288	0.253
tr-mmLDA	0.323	0.361	0.295	0.259
css-LDA	0.350	0.391	0.325	0.282
csGM-LDA	0.353	0.394	0.334	0.290

The proposed method achieves much more improvement on the ESP-Game dataset by comparing with a bit improvement on the Corel-5K dataset, demonstrating its effectiveness of modeling the large-scale dataset. For all the test sets, our proposed method performs best, and the instantiations of csGM-LDA and css-LDA clearly outperform others. We further compare the time required for training of different LDA models on ESP-Game and Corel-5K on a Intel Core i5-2410M CPU 2.30 GHz processor. Figure 7 gives the results. We observe that csGM-LDA is much faster than css-LDA. The reason is probably that the number of parameters in css-LDA is much larger than that of csGM-LDA. All the above results show the efficiency of our method.

6 Conclusions

In this paper, we propose a novel model, denoted as csGM-LDA, based on Laplacian regularized uncorrelated tensor representation for image annotation. The proposed annotation possesses two characteristics, namely: 1) images are represented by a set of uncorrelated tensorial descriptions and 2) class-specific information is integrated into semantic learning with the extension of the standard LDA model. The entire problem is formulated within the proposed framework, and csGM-LDA is presented to bridge the semantic gap between image contents and annotated tags. The experimental results demonstrate the effectiveness of our proposed method. Following the research on this line, we will further exploit region-based tensorial features for discriminative image representation and discuss the correlation of the class-specific information in a hierarchical LDA formalism.

Appendix

To maximize the lower bound in the E-Step that described in Section 4, we begin by expanding it with the factorizations of $P_{\Pi,Y,W,Z,V}(\theta, Y, W, Z, V|\alpha, \beta, \mu_{1:C}, \sigma_{1:C})$ and $q(\theta, Y, Z|\eta, \phi, \zeta)$:

$$\begin{aligned}&\mathcal{L}(\eta,\phi,\zeta;\alpha,\beta,\mu_{1:C},\sigma_{1:C}) = E_q[\log P_\Pi(\theta|\alpha)]\\&+E_q[\log P_{Y|\Pi}(Y|\theta)]+E_q[\log P_{W|Y}(W|Y,\beta)]\\&+E_q[\log P_Z(Z|\theta)]+E_q[\log P_{V|Z,W}(V|Z,W,\mu_{1:C},\sigma_{1:C})]\\&-E_q[\log q(\theta|\eta)]-E_q[\log q(Y|\phi)]-E_q[\log q(Z|\zeta)]\end{aligned}$$

We unfold each item, and obtain:

$$P_\Pi(\theta|\alpha) = \frac{\Gamma(\sum_k \alpha(k))}{\prod_k \Gamma(\alpha(k))}\prod_k \theta(k)^{\alpha(k)-1},$$

$$P_{Y|\Pi}(y_m|\theta) = \theta(y_m)$$

$$P_{Z|\Pi}(z_d|\theta) = \theta(z_d), P_{W|Y}(w_m|y_m,\beta) = \beta(y_m, w_m)$$

$$P_{V|Z,W}(v_d|z_d,\mu_{w_m},\sigma_{w_m}) = \frac{1}{\sqrt{2\pi}\sigma_{w_m}(z_d,d)} \times \exp\left(-\frac{(v_d-\mu_{w_m}(z_d,d))^2}{2\sigma^2_{w_m}(z_d,d)}\right)$$

To evaluate $\mathcal{L}(\eta,\phi,\zeta;\alpha,\beta,\mu_{1:C},\sigma_{1:C})$, we should measure $E_q(\log\theta)$. We find that the sufficient statistic in defining $P_\Pi(\theta|\alpha)$ is $\log\theta(k)$. Since the derivative of the

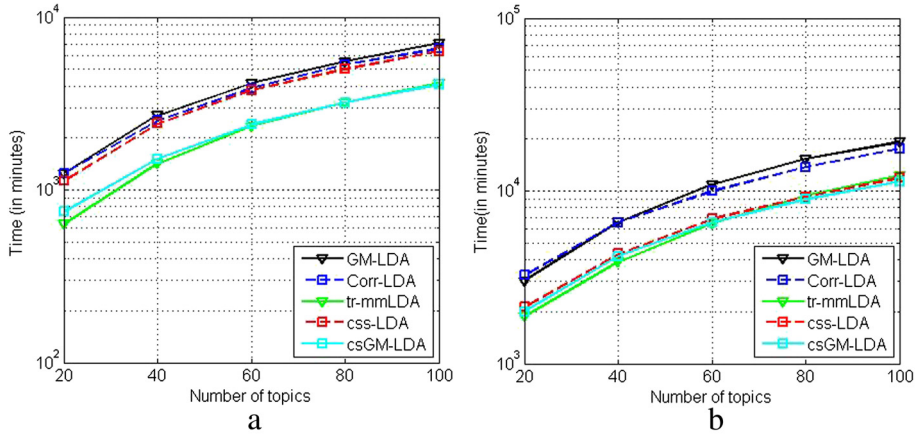

Figure 7 Time complexity of different LDA models on **(a)** Corel-5K and **(b)** ESP-Game.

log normalization factor with respect to the parameter is equal to the expectation of the sufficient statistic, we obtain:

$$E_q[\log\theta(k)] = \Psi(\eta(k)) - \Psi\left(\sum_{k=1}^{K}\eta(k)\right)$$

where $\Psi(\cdot)$ is the first derivative of $\log\Gamma(\cdot)$. Thus, we expand $\mathcal{L}(\eta,\phi,\zeta;\alpha,\beta,\mu_{1:C},\sigma_{1:C})$ as:

$$\begin{aligned}
&\mathcal{L}(\eta,\phi,\zeta;\alpha,\beta,\mu_{1:C},\sigma_{1:C}) = \\
&\textstyle\sum_{n=1}^{N}\left(\log\Gamma\left(\sum_{k=1}^{K}\alpha(k)\right) - \sum_{k=1}^{K}\log\Gamma(\alpha(k))\right. \\
&\left.+\sum_{k=1}^{K}(\alpha(k)-1)\,E[\log\theta_n(k)|\eta_n]\right) \\
&+\textstyle\sum_{n=1}^{N}\sum_{m=1}^{M}\sum_{k=1}^{K}\phi_{nm}(k)E[\log\theta_n(k)|\eta_n] \\
&+\textstyle\sum_{n=1}^{N}\sum_{m=1}^{M}\sum_{k=1}^{K}\phi_{nm}(k)\log\beta(k,w_{nm}) \\
&+\textstyle\sum_{n=1}^{N}\sum_{d=1}^{D}\sum_{k=1}^{K}\zeta_{nd}(k)E[\log\theta_n(k)|\eta_n] \\
&+\textstyle\sum_{n=1}^{N}\sum_{m=1}^{M}\sum_{d=1}^{D}\sum_{k=1}^{K}\zeta_{nd}(k)\log P_{V|Z,W}(v_{nd}|k,\mu_{w_{nm}},\sigma_{w_{nm}}) \\
&-\textstyle\sum_{n=1}^{N}\left(\log\Gamma(\sum_{k=1}^{K}\eta_n(k)) + \sum_{k=1}^{K}\log\Gamma(\eta_n(k))\right. \\
&\left.-\sum_{k=1}^{K}(\eta_n(k)-1)\,E[\log\theta_n(k)|\eta_n]\right) \\
&-\textstyle\sum_{n=1}^{N}\sum_{m=1}^{M}\sum_{k=1}^{K}\phi_{nm}(k)\log\phi_{nm}(k) \\
&-\textstyle\sum_{n=1}^{N}\sum_{d=1}^{D}\sum_{k=1}^{K}\zeta_{nd}(k)\log\zeta_{nd}(k)
\end{aligned}$$

In the E-Step, we first maximize the above function with respect to $\phi_{nm}(k)$. Observing that this is a constrained maximization since $\sum_{k=1}^{K}\phi_{nm}(k) = 1$, we first take the corresponding derivative:

$$\begin{aligned}
\frac{\partial\mathcal{L}}{\partial\phi_{nm}(k)} &= \Psi(\eta_n(k)) - \Psi\left(\sum_{j=1}^{K}\eta_n(j)\right) + \log\beta(k,w_{nm}) \\
&\quad - \log\phi_{nm}(k) - 1
\end{aligned}$$

Adding Lagrange multipliers to $\mathcal{L}$ and setting the derivative of the summation to zero yields the maximizing value of the variational parameter $\phi_{nm}(k)$:

$$\phi_{nm}(k) \propto \beta(k,w_{nm})\exp\left(\Psi(\eta_n(k)) - \Psi\left(\sum_{j=1}^{K}\eta_n(j)\right)\right)$$

Similarly, we can obtain the variational parameter $\zeta_{nd}(k)$ as follows:

$$\begin{aligned}
\zeta_{nd}(k) \propto &\left(\prod_{m=1}^{M}P_{V|Z,W}(v_{nd}|k,\mu_{w_{nm}},\sigma_{w_{nm}}\right) \\
&\times\exp\left(\Psi(\eta_n(k)) - \Psi\left(\sum_{j=1}^{K}\eta_n(j)\right)\right)
\end{aligned}$$

Next, we maximize $\mathcal{L}$ with respect to $\eta_n(k)$. Taking the corresponding derivative, we obtain:

$$\begin{aligned}
&\tfrac{\partial\mathcal{L}}{\partial\eta_n(k)} = \Psi'(\eta_n(k))\left(\alpha(k) - \eta_n(k)\right. \\
&\qquad\left.+\textstyle\sum_{m=1}^{M}\phi_{nm}(k) + \sum_{d=1}^{D}\zeta_{nd}(k)\right) \\
&-\textstyle\sum_{k=1}^{K}\Psi'\left(\sum_{j=1}^{K}\eta_n(j)\right)\left(\left(\alpha(k) - \eta_n(k)\right.\right. \\
&\left.\left.+\textstyle\sum_{m=1}^{M_n}\phi_{nm}(k) + \sum_{d=1}^{D}\zeta_{nd}(k)\right)\right)
\end{aligned}$$

Setting this equation to zero yields a maximum at:

$$\eta_n(k) = \alpha(k) + \sum_{m=1}^{M}\phi_{nm}(k) + \sum_{d=1}^{D}\zeta_{nd}(k)$$

In the M-Step, we also maximize the resulting lower bound $\mathcal{L}(\eta,\phi,\zeta;\alpha,\beta,\mu_{1:C},\sigma_{1:C})$ with respect to the model parameters $\{\alpha,\beta,\mu_{1:C},\sigma_{1:C}\}$. To estimate the Dirichlet parameter α, we first get the derivatives as:

$$\begin{aligned}
\frac{\partial\mathcal{L}}{\partial\alpha(k)} = \sum_{n=1}^{N}&\left(\Psi\left(\sum_{j=1}^{K}\alpha(j)\right) - \Psi(\alpha(k))\right. \\
&\left.+\left(\Psi(\eta_n(k)) - \Psi\left(\sum_{j=1}^{K}\eta_n(j)\right)\right)\right)
\end{aligned}$$

This derivative depends on $\alpha(j)$, for $j\neq k$, and we therefore must use an iterative method to find the maximal α. In particular, we can invoke the Newton-Raphson method with the update:

$$\begin{aligned}
&\alpha^{(i+1)}(k) = \alpha^{(i)}(k) - \\
&\left(\tfrac{\partial\mathcal{L}}{\partial\alpha^{(i)}(k)} - \frac{\sum_{j=1}^{K}\frac{\partial\mathcal{L}}{\partial\alpha^{(i)}(j)}\Big/\left(N\Psi'(\alpha^{(i)}(j))\right)}{1\Big/\left(N\Psi'\left(\sum_{j=1}^{K}\alpha^{(i)}(j)\right)\right) + \sum_{j=1}^{K}1\Big/\left(N\Psi'(\alpha^{(i)}(j))\right)}\right)\Big/ \\
&\left(N\Psi'(\alpha^{(i)}(k))\right)
\end{aligned}$$

To maximize $\mathcal{L}(\eta,\phi,\zeta;\alpha,\beta,\mu_{1:C},\sigma_{1:C})$ with respect to β, we add Lagrange multipliers and set the derivative of the summation to zero, we can obtain:

$$\beta(k,c) \propto \sum_{n=1}^{N}\sum_{m=1}^{M}\phi_{nm}(k)\delta(c,w_{nm})$$

As to $\{\mu_{1:C},\sigma_{1:C}\}$, we set the corresponding derivatives of $\mathcal{L}(\eta,\phi,\zeta;\alpha,\beta,\mu_{1:C},\sigma_{1:C})$ to zero and can exactly obtain:

$$\mu_c(k,d) = \frac{\sum_{n=1}^{N}\sum_{m=1}^{M}\zeta_{nd}(k)\delta(c,w_{nm})v_{nd}}{\sum_{n=1}^{N}\sum_{m=1}^{M}\zeta_{nd}(k)\delta(c,w_{nm})}$$

$$\sigma_c^2(k,d) = \frac{\sum_{n=1}^{N}\sum_{m=1}^{M}\zeta_{nd}(k)\delta(c,w_{nm})\,(v_{nd}-\mu_c(k,d))^2}{\sum_{n=1}^{N}\sum_{m=1}^{M}\zeta_{nd}(k)\delta(c,w_{nm})}$$

Competing interests

The authors declare that they have no competing interests.

Acknowledgement

This research was conducted with the support of National Natural Science Foundation of China (Grant No. 61271439).

Received: 26 November 2014 Accepted: 7 April 2015
Published online: 01 May 2015

References

1. H Ma, J Zhu, M-T Lyu, I King, Bridging the semantic gap between image contents and tags. IEEE Trans. Multimedia. **12**(5), 462–473 (2010)
2. J Liu, M Li, Q Liu, H Lu, S Ma, Image annotation via graph learning. Pattern Recognit. **42**, 218–228 (2009)
3. M Guillaumin, T Mensink, J Verbeek, C Schmid, in *Proc. IEEE Conf. Comput. Vis. Recognit*. Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation, (2009), pp. 309–316

4. N Zhou, W Cheung, G Qiu, X Xue, A hybrid probabilistic model for unified collaborative and content-based image tagging. IEEE Trans. Pattern Anal. Mach. Intell. **33**(7), 1281–1294 (2011)
5. S Yan, D Xu, Q Yang, L Zhang, X Tang, H-J Zhang, in *Proc. IEEE Conf. Comput. Vis. Recognit*. Discriminant analysis with tensor representation, (2005), pp. 526–532
6. Y Liu, Y Liu, S Zhong, K Chan, Tensor distance based multilinear globality preserving embedding: a unified tensor based dimensionality reduction framework for image and video classification. Expert Syst. Appl. **39**, 10500–10511 (2012)
7. Z Zhou, M Zhang, in *Proc. Adv. Neural Inf. Process. Syst*. Multi-instance multi-label learning with application to scene classification, (2006), pp. 1609–1616
8. B-K Bao, T Li, S Yan, Hidden-concept driven multilabel image annotation and label ranking. IEEE Trans. Multimedia. **14**(1), 199–210 (2012)
9. G Mesnil, A Bordes, J Weston, G Chechik, Y Bengio, Learning semantic representations of objects and their parts. Mach Learn. **94**(2), 281–301 (2014)
10. J Li, J Wang, Automatic linguistic indexing of pictures by a statistical modeling approach. IEEE Trans. Pattern Anal. Mach. Intell. **25**(9), 1075–1088 (2003)
11. Q Mao, I-H Tsang, S Gao, Objective-guided image annotation. IEEE Trans. Image Process. **22**(4), 1585–1597 (2013)
12. D Blei, A Ng, M Jordan, Latent dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
13. F Monay, D Gatica-Perez, Modeling semantic aspects for cross-media image indexing. IEEE Trans. Pattern Anal. Mach. Intell. **29**(10), 1802–1817 (2007)
14. D Blei, J McAuliffe, in *Proc. Adv. Neural Inf. Process. Syst*. Supervised topic models, (2008), pp. 121–128
15. Q Guo, N Li, Y Yang, G Wu, in *IEEE International Conference on Systems, Man, and Cybernetics*. Supervised LDA for image annotation, (2011), pp. 471–476
16. N Rasiwasia, N Vasconcelos, Latent Dirichlet allocation models for image classification. IEEE Trans. Pattern Anal. Mach. Intell. **35**(11), 2665–2679 (2013)
17. S Yan, D Xu, Q Yang, L Zhang, X Tang, H-J Zhang, Multilinear discriminant analysis for face recognition. IEEE Trans. Image Process. **16**(1), 212–220 (2007)
18. D Tao, X Li, X Wu, S Maybank, General tensor discriminant analysis and gabor features for gait recognition. IEEE Trans. Pattern Anal. Mach. Intell. **29**(10), 1700–1715 (2007)
19. J Yang, D Zhang, A Frangi, J Yang, Two-dimensional PCA: a new approach to appearance-based face representation and recognition. IEEE Trans. Pattern Anal. Mach. Intell. **26**(1), 131–137 (2004)
20. J Ye, R Janardan, Q Li, in *Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining*. Gpca: An efficient dimension reduction scheme for image compression and retrieval, (2004), pp. 354–363
21. H Lu, K Plataniotis, A Venetsanopoulos, MPCA: multilinear principal component analysis of tensor objects. IEEE Trans. Neural Netw. **19**(1), 18–39 (2008)
22. H Lu, K Plataniotis, A Venetsanopoulos, Uncorrelated multilinear principal component analysis for unsupervised multilinear subspace learning. IEEE Trans. Neural Netw. **20**(11), 1820–1836 (2009)
23. J Ye, Generalized low rank approximations of matrices. Mach. Learn. **16**(1), 167–191 (2005)
24. M Vasilescu, D Terzopoulos. Eur. Conf. Comput. Vis, (2002), pp. 447–460
25. D Xu, S Yan, L Zhang, S Lin, H-J Zhang, T Huang, Reconstruction and recognition of tensor-based objects with concurrent subspaces analysis. IEEE Trans. Circuits Syst. Video Technol. **18**(1), 36–47 (2008)
26. B Zhou, F Zhang, L Peng, Compact representation for dynamic texture video coding using tensor method. IEEE Trans. Circuits Syst. Video Technol. **23**(2), 280–288 (2013)
27. Z Li, Z Shi, X Liu, Z Shi, Modeling continuous visual features for semantic image annotation and retrieval. Pattern Recognit. Lett. **32**, 516–523 (2011)
28. S Nikolopoulos, S Zafeiriou, I Patras, I Kompatsiaris, High-order pLSA for indexing tagged images. Signal Process. **93**, 2212–2228 (2013)
29. D Blei, M Jordan, in *Proc. 26nd Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval*. Modeling annotated data, (2003), pp. 127–134
30. D Putthividhya, H Attias, S Nagarajan, in *Proc. IEEE Conf. Comput. Vis. Recognit*. Topic regression multi-modal latent Dirichlet allocation for image annotation, (2010), pp. 3408–3415
31. L Zhuang, H Gao, J Luo, Z Lin, Regularized semi-supervised latent Dirichlet allocation for visual concept learning. Neurocomputing. **119**, 26–32 (2013)
32. H Lu, K Plataniotis, A Venetsanopoulos, A survey of multilinear subspace learning for tensor data. Pattern Recognit. **44**, 1540–1551 (2011)
33. W-Y Ma, B Manjunath, Edgeflow: a technique for boundary detection and image segmentation. IEEE Trans. Image Process. **9**(8), 1375–1388 (2000)
34. H Liu, S Yan, in *International Conference on Machine Learning*. Robust graph mode seeking by graph shift, (2010), pp. 671–6783
35. K Murphy, *Machine Learning: A Probabilistic Perspective*. (The MIT Press, 2012)
36. H Müller, S Marchand-Maillet, T Pun, in *Proc. ACM Int'l Conf. Image and Video Retrieval*. The truth about Corel evaluation in image retrieval, (2002), pp. 38–49
37. L Ahn, L Dabbish, in *Proc. ACM SIGCHI Conf. on Human Factors in Computing Systems*. Labeling images with a computer game, (2004), pp. 354–363
38. T Ahonen, A Hadid, M Pietikainen, Face description with local binary patterns: application to face recognition. Mach. Learn. **28**(12), 2037–2041 (2006)